ESC-TR-92-142

Technical Report
973

Recognition-Time Speaker Adaptation in a Tied-Mixture HMM Continuous Speech Recognizer

B.F. Necioğlu
D.B. Paul

16 December 1996

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Prepared for the Defense Advanced Research Projects Agency under Air Force Contract F19628-95-C-0002.

Approved for public release; distribution is unlimited.

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-95-C-0002.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Gary Tutungian

Gary Tutungian
Administrative Contracting Officer
Contracted Support Management

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

RECOGNITION-TIME SPEAKER ADAPTATION IN A TIED-MIXTURE HMM CONTINUOUS SPEECH RECOGNIZER

B.F. NECIOĞLU and D.B. PAUL
Group 24

TECHNICAL REPORT 973

16 DECEMBER 1996

Approved for public release; distribution is unlimited.

LEXINGTON MASSACHUSETTS

ABSTRACT

All speech recognition systems, whether speaker-independent or speaker- dependent, require large amounts of training data to estimate the model parameters and, generally, the more training data available, the better the recognition performance. To improve the recognition performance of a system for a new speaker without having to train entirely new models, adapting the existing models during the recognition process is a practical solution. This report describes an investigation into the subject of recognition-time speaker adaptation of a tied-mixture HMM recognition system, with the goal of implementing a system which adapts to a new speaker during the course of its usage.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

Considering the large amounts of data required to train the models for a *speaker-dependent* (SD) or *speaker-independent* (SI) continuous speech recognizer (CSR), adaptation of the existing models to a new speaker using relatively small amounts of data can be a desirable alternative.

In the *rapid enrollment* approach to adaptation, a small amount of data is collected from the new speaker to modify a prior set of model parameters and these new parameters are used to recognize the incoming utterances [3,7,8]. In this report, we present the results of a study whose goal is to achieve a recognition system which starts with a set of speaker-dependent or speaker-independent initial acoustic models and continuously adapts itself as it is being used by the current speaker, in an either supervised or unsupervised fashion.

The *supervised* adaptation scheme requires that the recognition output be corrected by the speaker if it is erroneous, while the *unsupervised* version simply uses the uncorrected recognizer output for adaptation.

The recognition system used to investigate this adaptation problem is the Lincoln large-vocabulary continuous speech recognizer, which is a *stack-decoder* based system with tied-mixture Gaussian Hidden Markov Model (HMM) acoustic models [4,5,6]. In particular, the acoustic models used in the recognizer for this work are non-cross-word semi-phone based, and provide a simpler framework for the initial research into the subject, compared to the more complex cross-word triphone based models that can also be used by the recognizer. [4,5].

2. RECOGNITION-TIME SPEAKER ADAPTATION

Using the recognized utterances instead of a pre-determined static set of spoken sentences to modify the recognizer models is what distinguishes recognition-time adaptation from rapid enrollment. The recognition-time adaptation is performed progressively on a sentence-by-sentence basis and therefore rapid enrollment, by accessing all the data at once, has the advantage of being a block process when the models are modified. Using a limited amount of data to modify the large number of parameters of the acoustic models requires robust estimation techniques and/or modification of only a selected subset of the parameters. On the other hand, rapid enrollment, which requires the user to record a set of sentences beforehand, lacks the ease of use of recognition-time adaptation.

The Lincoln HMM CSR used for this research on recognition-time adaptation utilizes *tied-mixture* Gaussian semi-phone acoustic models. The observation probability at a given state s is given by a Gaussian tied-mixture density with M components:

$$P_s(\mathbf{o}_t) = \sum_{i=1}^{M} c_{is} G_i(\mathbf{o}_t), \qquad (1)$$

where $\mathbf{o}_t$ is the observed frame vector at time t, $G_i(\mathbf{o}_t) = N(\mu_i, \Sigma_i)$ and $\sum_{i=1}^{M} c_{is} = 1$. With the assumption of independence between the components of the observation vectors, Σ_i becomes diagonal. The set of acoustic model parameters to be estimated during the training phase for the Lincoln HMM CSR consists of the Gaussian means (μ_i) and grand variances ($\Sigma = \Sigma_i\,;\ \forall i$), which are tied across all the acoustic models, and the mixture weights (c_{is}) and state transition probabilities.

As recognition-time speaker adaptation requires parameter modification using small amounts of data, adapting all of the acoustic model parameters was not attempted. The tied-mixture Gaussian density framework offers a nice formalism for adaptation because during training, each and every observation contributes to all of the Gaussians in the mixture density, enabling robust estimation from small amounts of data. The initial investigations focused on the adaptation of only the Gaussian means of the existing set of tied-mixture semi-phone models, while keeping the grand Gaussian variances, mixture weights and other HMM parameters fixed.

After recognition of each sentence, the Gaussian means are modified according to the observation data. Since the amount of data from the recognized sentence is very small, new estimates must be smoothed by the previous values. Two methods for combining the initial/past models with the new data have been explored: (1) infinite memory, and (2) exponentially dying memory [2]. In both cases, the Viterbi alignment of the observations to the recognized sentence is used to make a new estimate of the Gaussian means with the Estimate-Maximize (EM) algorithm:

- Estimate:

$$\gamma_{it} = \frac{c_{is_t} G_i(\mathbf{o}_t)}{\sum_{i=0}^{M} c_{is_t} G_i(\mathbf{o}_t)}$$

- Maximize:

$$\mu_{i_{\mathsf{new}}} = \frac{\sum_t \gamma_{it} \mathbf{o}_t}{\sum_t \gamma_{it}},$$

where the state sequence s_t is given by the Viterbi alignment of the observations $\mathbf{o}_t$. Since a non-tied mixture Gaussian can be viewed as a special case of a tied-mixture Gaussian, these reestimation equations also apply to non-tied mixtures.

For the infinite memory case, to obtain the adapted mean, the new estimate of the mean ($\mu_{i_{\mathsf{new}}}$) is combined with the old mean as if the new data were pooled with the previous training data:

$$\mu_{i_{\mathsf{adapted}}} = \frac{\mu_{i_{\mathsf{old}}} \sum_{t \in \mathsf{old}} \gamma_{it} + \mu_{i_{\mathsf{new}}} \sum_{t \in \mathsf{new}} \gamma_{it}}{\sum_{t \in \mathsf{old}} \gamma_{it} + \sum_{t \in \mathsf{new}} \gamma_{it}} \tag{2}$$

Upon completion of adaptation, the next sentence is recognized with the new parameter set and the same procedure is repeated. This has the effect of assigning less weight to the newly arriving data as adaptation continues, which slows down the adaptation speed. The exponentially dying memory case, keeps an exponentially decaying memory of the original models instead of an infinite one. This is achieved by linearly combining the existing Gaussian means with the means estimated from the observations of the newly recognized sentence:

$$\mu_{i_{\mathsf{adapted}}} = (1-\lambda)\,\mu_{i_{\mathsf{old}}} + \lambda\,\mu_{i_{\mathsf{new}}} \tag{3}$$

By fixing the adaptation weight λ to a constant value, the contribution of the initial speaker-dependent or speaker-independent model to the adapted model is made to decay exponentially, which also gives a means of control over the speed of adaptation. The infinite memory case can be obtained by assigning

$$\lambda = \frac{\sum_{t \in \mathsf{new}} \gamma_{it}}{\sum_{t \in \mathsf{old}} \gamma_{it} + \sum_{t \in \mathsf{new}} \gamma_{it}}, \tag{4}$$

instead of keeping it constant. For $\lambda = 0.1$, the contribution of the initial non-adapted model to the adapted one becomes about 0.5% by the time 50 sentences are recognized. Setting λ higher results in faster adaptation but since it results in noisier estimates of the parameters, the limit of the recognition performance may be poorer in the long run, whereas a smaller λ gives a slower speed of adaptation, but may have a better performance limit. However, given that the purpose of adaptation is to improve performance with minimal effort, λ could be set to be large at the beginning and made to decrease with the number of sentences that have been recognized so far, to achieve speed in the short term and have a better performance limit in the future.

3. EXPERIMENTS AND RESULTS

The adaptation experiments are performed using the RM1 and RM2 sections of the DARPA Resource Management corpus. For cross-speaker adaptation tests, speaker-dependent models trained using two (one male, one female) of the four speakers of the RM2 database (2400 sentences per speaker) are adapted to the twelve RM1 test speakers and for speaker-independent adaptation tests, an RM1 speaker-independent (SI-109) model and a modified version is adapted to the same twelve RM1 test speakers. All the starting SD and SI models were trained with speaker and channel equalization, which was performed by subtracting the sentence average speech-only mel-cepstral observation vector from every mel-cepstral observation vector across the sentence. This can be viewed as a "dc" removal, resembling a portion of the RASTA methodology [1].

The available development test set for the SD-RM1 speakers consist of 100 sentences per speaker, and the results in Table 1 reflect the average word error (substitutions + insertions + deletions) rate over all twelve speakers, with their respective standard deviations. To observe the course of the continuing adaptation process, the set of 100 test sentences per speaker was divided into 25 sentence portions, and scored within each 25 sentence portion.

As expected, supervised recognition-time adaptation results show that the exponentially decaying memory achieves much better results than the infinite memory version. Several different λ values were tested to find an optimum adaptation rate, and for the span of the available 100 test sentences, λ=0.1 was decided to give the best trade-off between adaptation speed and long-term performance limit (Figure 1 and Figure 2). A smaller weight (λ=0.04) was slower to adapt the system, while a larger one (λ=0.2), although adapting faster, had a poorer recognition performance limit for almost all of the configurations. Actually, the self-adaptation experiments, in which the speakers' own models are used to adapt to their own utterances, show that the higher the adaptation rate the worse the long-term performance limit.

The exponentially decaying memory adaptation is demonstrating a much better percentage performance improvement for the cross-speaker cases than it does for the speaker-independent case (Table 1 and Figures 1 and 2). Particularly, after adapting to 100 sentences (approximately 5 minutes of speech), the cross-speaker/cross-gender case gets down to 15.6% word error rate (83.4% reduction in error) and the cross-speaker/same-gender case gets down to 9.5% word error rate (59.1% reduction), with respect to their control experiments, while the speaker-independent case gets down to 7.4% word error rate (36.8% reduction). Since the large set of mixture weights is kept fixed and only the Gaussian means are adapted, it is likely that the sharper distributions of the speaker-dependent models converge more rapidly to the new speaker's, while broad distributions of the speaker-independent models are slow to do so. The modified SI-109 model is trained in two phases [4]: during the first phase, estimation of SI mixture weights using a set of *speaker-dependent* Gaussian means is performed, and in the second phase, mixture weights are fixed and a single set of *speaker-independent* Gaussian parameters are trained. Because of the way the mixture weights are trained, this strategy is intended to give a "sharper" set of weights for the SI model (more like an SD model), which becomes more suitable to Gaussian-parameter-only speaker-adaptation. This is reflected in the results, and although the modified SI-109 starts with a poorer performance than the standard SI-109, after adapting to 100 sentences, it beats the other and goes down to 6.2% word error rate, achieving 55.7% reduction in error with respect to its non-adapted version (Table 1).

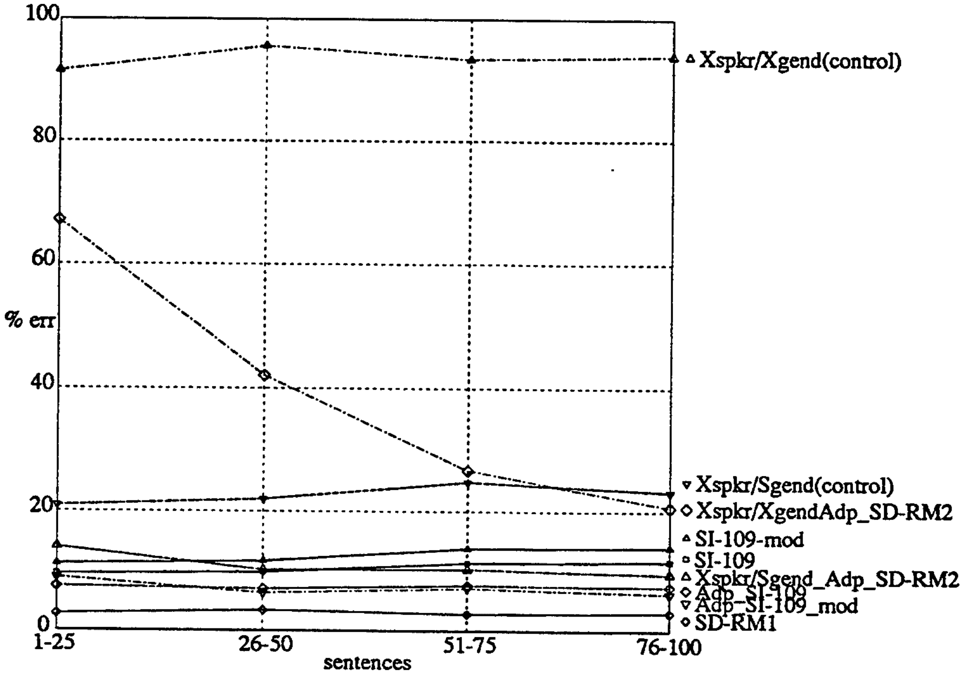

Figure 1. Word error rates from Table 1, for the control and λ=.1 adaptation conditions. (Low error rate portion–below 15%– detailed in Figure 2).

Using the optimum weight (λ=0.1), the same set of experiments were conducted in an unsupervised fashion. The results obtained are very promising (Table 1 and Figure 3), suggesting that the system can adapt to a new user without any corrections by the user at the recognition stage. For all but cross-speaker/cross-gender adaptation experiments, the performance improvement difference between the supervised and unsupervised cases is statistically insignificant. But still, after adapting to 100 test sentences, cross-speaker/cross-gender configuration achieves the best percentage performance improvement among all the other schemes for both the supervised and unsupervised cases. In this configuration, from a non-adapted word error rate of 94.0%, supervised adaptation gets down to 20.7% (78.0% reduction) and the unsupervised case gets down to 30.1% (68.0% reduction) word error rate, for the last 25 test sentence portion. Even though our experiments show that a system which has more than 90% word error rate at the starting point

TABLE 1

Word error rate (%) with standard deviations, averaged over SD-RM1 speakers, within each 25 sentence portion. (Error reduction figures with respect to the relevant control case, for the last 25 sentence portions.)

Experiments	Recognized Sentences from RM1 / Development Test Set				Red. (%)
	1-25	26-50	51-75	76-100	
SI-109	9.5(0.6)	9.8(0.6)	11.4(0.6)	11.7(0.6)	-
Adapted SI-109 (infinite memory)	9.5(0.6)	9.8(0.6)	11.4(0.6)	11.7(0.6)	0
Adapted SI-109 (λ = .04)	8.2(0.6)	7.6(0.5)	7.8(0.5)	8.6(0.6)	26.5
Adapted SI-109 (λ = .1)	7.4(0.5)	7.0(0.5)	7.6(0.5)	7.4(0.5)	36.8
Adapted SI-109 (λ = .2)	7.7(0.5)	6.7(0.5)	8.2(0.5)	7.8(0.5)	33.3
Unsupervised adapted SI-109 (λ=.1)	7.3(0.5)	7.1(0.5)	7.3(0.5)	7.7(0.5)	34.2
Modified SI-109	11.2(0.6)	11.7(0.6)	13.8(0.7)	14.0(0.7)	-
Adapted modified SI-109 (inf. mem.)	11.2(0.6)	11.5(0.6)	13.4(0.7)	13.8(0.7)	1.4
Adapted modified SI-109 (λ=.04)	9.9(0.6)	7.5(0.5)	7.8(0.5)	7.5(0.5)	46.4
Adapted modified SI-109 (λ=.1)	8.9(0.6)	6.3(0.5)	7.1(0.5)	6.2(0.5)	55.7
Adapted modified SI-109 (λ=.2)	8.0(0.6)	6.2(0.5)	7.0(0.5)	6.8(0.5)	51.4
Unsup. adapted modified SI-109 (λ=.1)	9.1(0.6)	6.5(0.5)	6.9(0.5)	6.7(0.5)	52.1
SD-RM1	2.8(0.3)	3.4(0.4)	2.8(0.3)	3.0(0.3)	-
Self-adapted SD-RM1 (inf. mem.)	2.8(0.3)	3.5(0.4)	2.8(0.3)	3.1(0.3)	-3.3
Self-adapted SD-RM1 (λ=.04)	3.2(0.4)	3.3(0.3)	3.1(0.3)	3.2(0.3)	-6.7
Self-adapted SD-RM1 (λ=.1)	3.3(0.4)	4.0(0.4)	3.8(0.4)	3.8(0.4)	-26.7
Self-adapted SD-RM1 (λ=.2)	4.0(0.4)	4.1(0.4)	4.3(0.4)	4.0(0.4)	-33.3
X-spkr/Same-gender (control)	20.9(0.8)	22.0(0.8)	24.8(0.8)	23.2(0.8)	-
X-spkr/S-gend adp. SD-RM2 (inf. mem.)	20.6(0.8)	21.7(0.8)	24.2(0.8)	22.1(0.8)	4.7
X-spkr/S-gend adp. SD-RM2 (λ=.04)	17.2(0.8)	11.9(0.6)	12.3(0.6)	10.7(0.6)	53.9
X-spkr/S-gend adp. SD-RM2 (λ=.1)	14.0(0.7)	10.2(0.6)	10.3(0.6)	9.5(0.6)	59.1
X-spkr/S-gend adp. SD-RM2 (λ=.2)	13.0(0.7)	10.3(0.6)	11.3(0.6)	10.7(0.6)	53.9
U. X-spkr/S-gend adp. SD-RM2 (λ=.1)	14.1(0.7)	10.9(0.6)	11.1(0.6)	10.5(0.6)	54.7
X-spkr/X-gender (control)	91.6(0.6)	95.6(0.4)	93.5(0.5)	94.0(0.5)	-
X-spkr/X-gend adp. SD-RM2 (inf. mem.)	91.4(0.6)	95.8(0.4)	91.4(0.5)	93.3(0.5)	.7
X-spkr/X-gend adp. SD-RM2 (λ=.04)	81.0(0.8)	65.8(0.9)	49.6(1.0)	39.2(1.0)	58.3
X-spkr/X-gend adp. SD-RM2 (λ=.1)	67.1(1.0)	42.1(1.0)	26.7(0.9)	20.7(0.8)	78.0
X-spkr/X-gend adp. SD-RM2 (λ=.2)	55.8(1.0)	26.9(0.9)	19.7(0.8)	15.6(0.7)	83.4
U. X-spkr/X-gend adp. SD-RM2 (λ=.1)	74.1(0.9)	51.9(1.0)	35.9(0.9)	30.1(0.9)	68.0

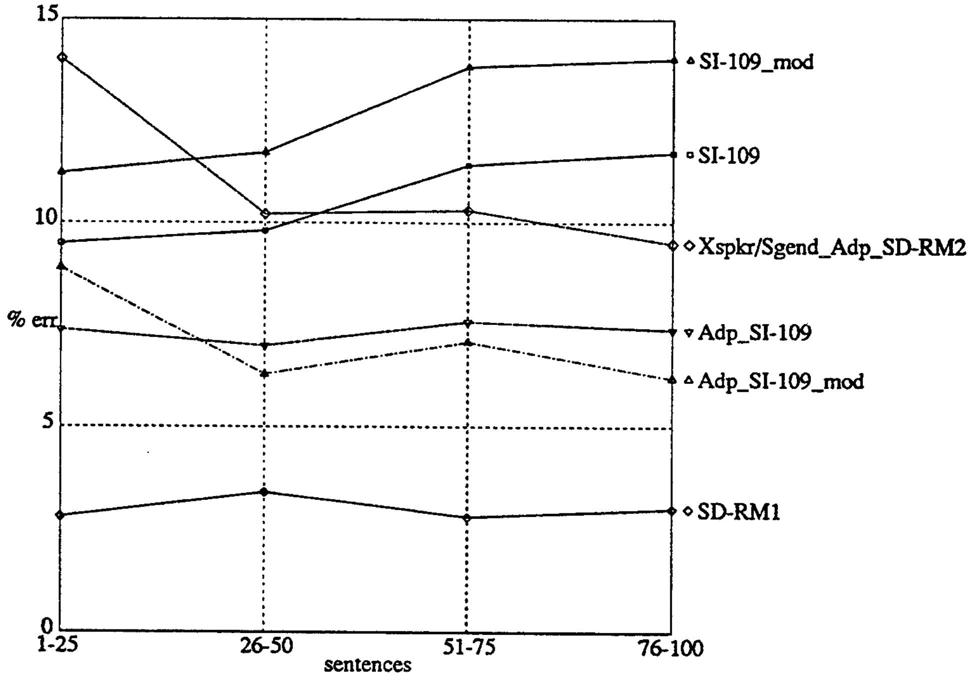

Figure 2. Word error rates from Table 1, for the control and λ=.1 adaptation conditions (detail of Figure 1).

of adaptation (cross-speaker/cross-gender case) can still get a dramatic improvement with unsupervised adaptation using just 100 sentences, this fact does not imply that convergence is always guaranteed. Contrary to the supervised case, an unsupervised adaptation procedure *might* diverge if the starting word error rate is too high.

All the aforementioned experiments were conducted using non-cross-word semiphone models with two observation streams (cepstra and difference cepstra), and although this was convenient for the initial research and development phase, the results do not represent the best performance achievable with the current Lincoln HMM CSR. Additional SI recognition-time adaptation experiments were conducted using cross-word triphone models with three observation streams (cepstra, difference cepstra, and second difference cepstra), for both the supervised and unsupervised cases, and the results show that recognition-time speaker adaptation improves performance for the best system as well, with 49.9% word error reduction over the non-adapted SI-109 for the supervised case and 45.9% for the unsupervised case, after an adaptation run over 100 sentences for the twelve RM1 speakers (Table 2 and Figure 4).

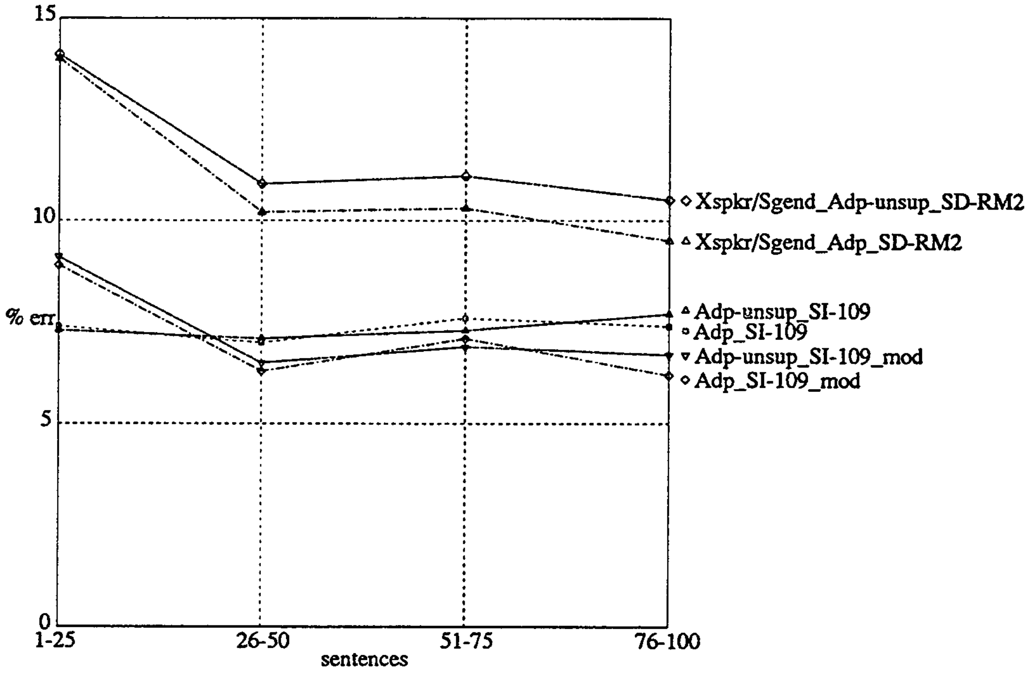

Figure 3. Word error rates from Table 1. Comparison of the supervised and unsupervised adaptation conditions, with λ=.1.

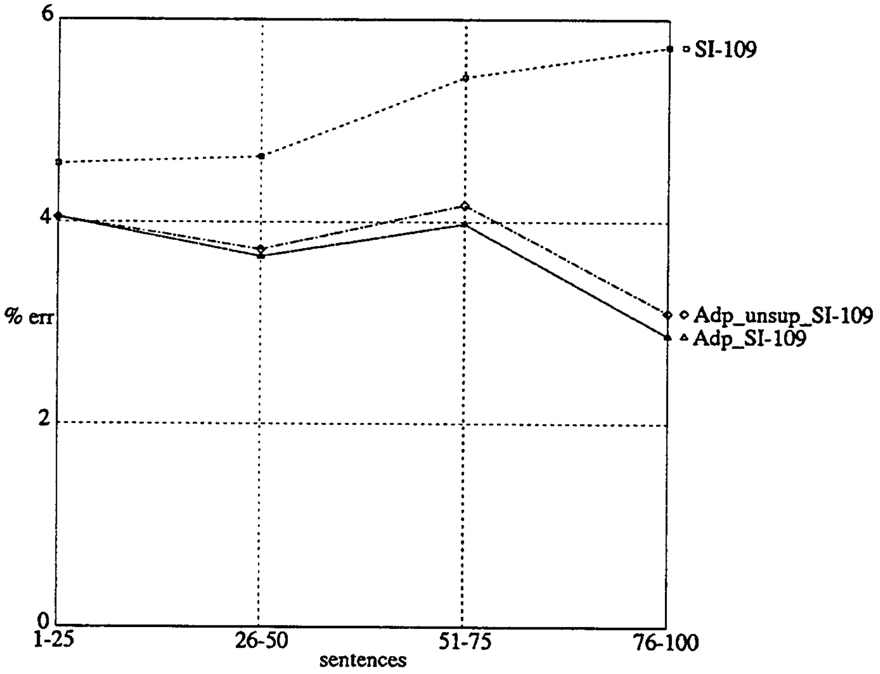

Figure 4. Word error rates from Table 2, for the control and λ=.1 adaptation conditions.

TABLE 2

Adaptation experiment results for the best configuration of the Lincoln tied-mixture HMM CSR (cross-word triphone models, 3 observation streams). Word error rate (%) with standard deviations, averaged over SD-RM1 speakers, within each 25 sentence portion. (Error reduction figures with respect to the control case, for the last 25 sentence portions.).

Experiments	**Recognized Sentences from RM1 / Development Test Set**				**Red. (%)**
	1-25	26-50	51-75	76-100	
SI-109 (control)	4.58(.42)	4.65(.41)	5.43(.44)	5.73(.46)	-
Adapted SI-109 (λ=.1)	4.05(.40)	3.66(.37)	3.98(.38)	2.87(.33)	49.9
Unsupervised adapted SI-109 (λ=.1)	4.05(.40)	3.73(.37)	4.17(.39)	3.10(.34)	45.9

4. SUMMARY AND FUTURE DIRECTIONS

This research into the supervised/unsupervised recognition-time speaker adaptation of a tied-mixture Gaussian CSR confirms the viability of the concept. Infinite training memory and exponentially decaying training memory approaches to the adaptation of the means of Gaussian mixture components have been investigated, and significant performance reductions have been achieved, both with the research configured Lincoln HMM CSR, and its best performing configuration.

Still a number of aspects of the problem remain to be investigated. The exponentially decaying memory approach, although achieving very good performance improvements, assigns the same weight to the new estimates regardless of the amount of data used to reestimate each Gaussian. Actually, the adaptation weight λ should be dynamic rather than static, and vary in proportion to the amount of data used to reestimate each Gaussian. The rate should also be set high in the beginning to have a fast adaptation, and decrease with the total time spent by each speaker using the system to improve the recognition performance limit over the course of the adaptation.

Although this study has examined adaptation of the means of Gaussian mixture components only, the joint adaptation of means, variances and tied-mixture weights should be explored as well.

REFERENCES

1. H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "RASTA-PLP Speech Analysis Technique," *Proc. ICASSP'92*, Vol. I, pp. 121–124, San Francisco, CA, March 1992.
2. E. A. Martin, R. P. Lippmann, and D. B. Paul, "Dynamic Adaptation of Hidden Markov Models for Robust Isolated-Word Speech Recognition," *Proc. ICASSP'88*, pp. 52–54, New York, NY, April 1988.
3. B.F. Necioğlu, M. Ostendorf, J.R. Rohlicek, "A Bayesian Approach to Speaker Adaptation for the Stochastic Segment Model," *Proc. ICASSP'92*, Vol. I, pp. 437–440, San Francisco, CA, March 1992.
4. D. B. Paul, "New Results with the Lincoln Tied-Mixture HMM CSR System," *Proc. 1991 DARPA Speech and Natural Language Workshop*, pp. 65–70, Pacific Grove, CA, February 1991.
5. D. B. Paul, "The Lincoln Tied-Mixture HMM Continuous Speech Recognizer," *Proc. ICASSP'91*, pp. 329–332, Toronto, Canada, May 1991.
6. D. B. Paul, "An Efficient A* Stack Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model," *Proc. ICASSP'92*, Vol. I, pp. 25–28, San Francisco, CA, March 1992.
7. D. Rtischev, "Speaker Adaptation in a Large-Vocabulary Speech Recognition System," M.S. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.
8. R. Schwartz, Y.L. Chow, and F. Kubala, "Rapid Speaker Adaptation Using a Probabilistic Spectral Mapping," *Proc. ICASSP'87*, pp. 633–636, Dallas, TX, April 1987.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE 16 December 1996	3. REPORT TYPE AND DATES COVERED Technical Report

4. TITLE AND SUBTITLE

Recognition-Time Speaker Adaptation in a Tied-Mixture
HMM Continuous Seech Recognizer

5. FUNDING NUMBERS

C — F19628-95-C-0002
PE — 62702E, 62301E, 61101E
PR — 337

6. AUTHOR(S)

B.F. Necioğlu and D.B. Paul

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
244 Wood Street
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-973

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DARPA/ITO
3701 N. Fairfax Dr.
Arlington, VA 22203-1714

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ESC-TR-92-142

11. SUPPLEMENTARY NOTES

Approved for public release; distribution is unlimited.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

All speech recognition systems, whether speaker-independent or speaker-dependent, require large amounts of training data to estimate the model parameters and, generally, the more training data available, the better the recognition performance. To improve the recognition performance of a system for a new speaker without having to train entirely new models, adapting the existing models during the recognition process is a practical solution. This report describes an investigation into the subject of recognition-time speaker adaptation of a tied-mixture HMM recognition system, with the goal of implementing a system which adapts to a new speaker during the course of its usage.

14. SUBJECT TERMS

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT